Generic and domain-specific aspects of the Waxholm NLP and dialog modules

Rolf Carlson and Sheri Hunnicutt

Department of Speech, Music and Hearing,
KTH, Sweden

ABSTRACT

In this paper we give an overview of the NLP and dialog component in the Waxholm spoken dialog system. We will discuss how the dialog and the natural language component are modeled from a generic and a domain-specific point of view. Dialog management based on grammar rules and lexical semantic features is implemented in our parser. The notation to describe the syntactic rules has been expanded to cover some of our special needs to model the dialog. The parser is running with two different time scales corresponding to the words in each utterance and to the turns in the dialog. Topic selection is accomplished based on probabilities calculated from user initiatives. Results from parser performance and topic prediction are included.

1. Background

Our research group at KTH has, for some years, been building a generic system in which speech synthesis and speech recognition can be studied in a man-machine dialog framework. The demonstrator application, Waxholm, gives information on boat traffic in the Stockholm archipelago. It references time tables for a fleet of some twenty boats from the Waxholm company which connects about two hundred ports. The system has been presented on several occasions, for example [1,2,3,4].

In our contribution we will describe in some detail the current effort to model the NLP and dialog processing in the Waxholm spoken dialog system. Our objective is to develop a dialog management module which can handle the type of interaction that can occur in our chosen domain. The Waxholm system should allow user initiatives, without any specific instructions to the user, complemented by system questions to achieve the user's goal. We will discuss how the dialog and the natural language component are modeled from a generic and a domain-specific point of view.

2. Natural language modeling

In this section we will give a short review of the natural language component, STINA. Some of our fundamental concepts were initially inspired by TINA, a parser developed at MIT, [5]. STINA is knowledge based and contains a context-free grammar which is compiled into an ATN. (A detailed description of the parser can be found in [3].) Probabilities are assigned to each arc after training. These probabilities are primarily used to reduce search time and for hypothesis pruning. The parsing is done in three steps. The first step makes use of broad categories such as nouns, while the following step expands these into more detailed solutions. The last step involves recalculation of hypothesis probabilities according to a multi-level N-gram model.

2.1. Domain dependent feature system

The feature system used in the parser plays an important role. Each lexical entry can have domain specific semantic features associated to it in addition to the basic syntactic features. The semantic features as used in STINA can be divided into two different classes, basic semantic features and function features. Basic features such as BOAT and PORT give a simple description of the semantic property of a word and are often domain specific. These features are hierarchically structured. In our domain we have specified that a PORT is part of an ISLAND, which is part of a REGION, which is part of a PLACE, which is part of the WORLD.

The second type of semantic features is the "function features." These features are not hierarchical. Typically they are associated with an action, such as TO_PLACE indicating the destination in an utterance regarding travel (Example 1). The function features are also node names in the parser. A verb can have function features set, allowing or disallowing a certain type of modifier to be part of a clause. The action itself in the TO_PLACE example has, of course, a broader scope than the traveling domain, and includes movements between any reference points. Thus, the node TO_PLACE is specified as a prepositional phrase starting with "to" and followed by any nominal expression. The scope of the phrase is changed according to the domain by training.

Example 1: (TO_PLACE ("to"/TO "Waxholm"/noun))

The function features are powerful tools to control the analysis of responses to questions from the dialog module. The question "Where do you want to go?" conditions the parser to accept a simple port name or a prepositional phrase including a port name as a possible response from the user. This property of STINA gives the parser some of the advantages of a functional grammar parser.

Terminal node evaluation is primarily carried out on the grammatical features. If this basic constraint evaluation is accepted, the semantic features are also evaluated. The hierarchical structure has importance for the rule writing. During the unification process all semantic features which belong to the same semantic branch in the feature tree are considered. The whole tree of the lexical entry is moved into the hypothesis including the leaves on the feature tree. In our traveling domain a port name will keep its PORT feature even if only the PLACE is noted in the grammar. This has several advantages. The rules or terminal specifications do not have to be more specific than necessary and the domain knowledge can, to some extent, be part

of the lexicon rather than the rules. This mechanism is extensively used in the sublanguage grammar for our application. In the next section we will see how the introduction of domain dependent terminal nodes is delayed during the parsing process.

2.2. General grammar to subgrammar

It has been an ambition in our work to create a general grammar which at least covers the type of dialog found in our domain. After an utterance initially has been parsed, we have a hypothesis in terms of grammar nodes and generic terminals such as nouns. In the next step the terminals are replaced by more domain specific labels. A domain specific list of possible terminals is processed by the parser during initialization and each such terminal is associated a generic terminal node. The domain specific nodes are typically constrained by domain specific semantic features in addition to the basic syntactic ones. In our case we have defined a number of terminals, such as port, hotel, boat and time-table. These are all part of the noun class and will replace the "noun" terminal whenever appropriate according to the semantic features of the lexical entry. In our application, then, the lexicon defines that there are nouns with a specific semantic feature, PORT, and is able to separate them from other nouns. The simple phrase in Example 1 is turned into the phrase "TO port" since Waxholm is a port and the terminal port is part of the noun class, (see Example 2.)

Example 2: (TO_PLACE ("to"/TO "Waxholm"/port))

With this approach we can formulate a general grammar and make it domain specific with the help of the feature system and lexical specifications.

The described method has some attractive side effects. Since the network specified by the ATN has generic terminals, the number of nodes and transitions are less than if the grammar were more specific. This makes the parsing faster since fewer hypotheses have to be evaluated. However, the probability calculation is less informative based on broad categories and has to be reconsidered. In our case this is done with the help of N-gram models.

2.3. N-gram models

It seems to be a general consensus that N-gram models, in the context of speech understanding, have at least as good predictive power as regular knowledge based grammars [6,7]. However, some research, such as the work by Seneff et al. [8], has shown, that a knowledge based parser including multilayered probabilities has some advantages. This is specially true for the following processing in the dialog system.

In STINA, smoothed N-gram models are used in addition to the regular transition probabilities. N-gram probabilities are added to the node probabilities after the domain specific node replacements have been performed and before a hypothesis is pushed on the probability ordered N-best stack. The N-gram probabilities include not only terminal node sequences but also phrase level heads. The work by Moore et al. [9] has earlier shown the advantage in adding phrase heads in the N-gram calculation. In Example 3 the hypothesis score calculation includes for example: p(boat | "TOP+SUBJ"), p(v | "SUBJ+boat+VP"), p(TO | "VP+v+TO_PLACE") and p(port | "VP+v+TO_PLACE"). We have expanded the calculation to also include phrase level head node probabilities. However, they are based on phrase level head sequencies. p(SUBJ | "TOP"), p(VP | "TOP+SUBJ"), and p(TO_PLACE | "TOP+SUBJ+VP").

Example 3: (TOP (SUBJ "båten"/boat)
(VP "går"/v (TO_PLACE ("till"/TO "Vaxholm"/port))))

As an additional example we find that the utterance "I want to go from X to Y" is more probable in our application than "I want to go to X from Y" as reflected in the node N-gram probabilities. Thus, this last step of hypothesis scoring is a powerful method to adjust the general grammar to the domain specific analysis that is needed. Certain phrases and phrase sequences will be well described in the N-gram statistics.

3. Dialog modeling

Two major ideas have been guiding the work on the dialog model. First, the dialog should be described by a grammar. Second, the dialog should be probabilistic. In our system, dialog building blocks are described by nodes. Each node has specifications concerning, for example, dialog action, constraint evaluation and system response. A graphical interface to the system presents the dialog grammar graphically. Both the syntax and the dialog can be modeled and edited graphically with this tool.

Topic selection is accomplished based on probabilities calculated from user initiatives [3,4]. Lexical semantic information combined with semantic grammar nodes are used as factors in this calculation. The topic selection based on probabilities in our system has similarities with the effort at AT&T [10]. A special session in the Eurospeech 1995 conference was devoted to word spotting including topic spotting based on keywords. The work by Nowel and Moore [11] goes one step further exploring non-word based topic spotting.

A modification of the domain implies an addition of how to handle a new topic, but it is our ambition that the implementation and the training procedures should, as much as possible, be kept the same.

3.1. Topic selection

The decision about which topic path to follow in the dialog is based on several factors such as the dialog history and the content of the specific utterance. The utterance is coded in the form of a "semantic frame" with slots corresponding to both the grammatical analysis and the specific application. The structure of the semantic frame is automatically created based on the rule system.

Each semantic feature found in the syntactic and semantic analysis is considered in the form of a conditional probability to decide on the topic. The probability for each topic is expressed as: p(topic|F), where F is a feature vector including all semantic features used in the utterance. Thus, the BOAT feature can be a

strong indication for the TIME_TABLE topic but this can be contradicted by a HOTEL feature. The topic prediction has been trained using a labeled set of utterances taken from the Waxholm database. Only utterances indicating a topic (about 1200) have been included in this set. The probability is calculated according to: p = (n+1)/(N+2), where N = number of times a feature can be a terminal node in the feature tree, and n = number of times a feature actually is a terminal node in a topic indicating utterance.

3.2. Introduction of a new topic

In this section we will give a simple example of how a new topic can be introduced. Suppose we want to create a topic called "out of domain." First a topic node is introduced in the rule system. Some new words probably need to be included in the lexicon and labeled with a semantic feature showing that the system does not know how to deal with the subjects these words relate to. Then a synthesis node might be added with an output informing the user about the situation. Example sentences must be created that illustrate the problem and the dialog parser must be trained with these sentences labeled with the "out of domain" topic. Since the topic selection is done by a probabilistic approach that needs application-specific training, data collection is of great importance for the progress of the project.

4. Evaluation of the NLP and dialog moduls

Evaluation of the system has been performed using part of the Waxholm database. In this database, speech and text data was collected using the Waxholm system. Initially, a "Wizard of Oz" replaced the speech recognition module. A full report on the data collection and data analysis can be found in [1].

The database was collected using preliminary versions of each module in the Waxholm system. This procedure has advantages and disadvantages for the contents of the database. System limitations will already from the beginning put constraints on the dialog, making it representative for a human-machine interaction. However, since the system was under development during the data collection, it was influenced by the system status at each recording time. After about half of the recording sessions, the system was reasonably stable, and the number of system "misunderstandings" had been reduced. In this section, we will discuss parser performance and topic selection. As research on dialog systems develops, it becomes more important to develop new methods to evaluate human-machine interaction.

4.1. Test material

The test material used in the experiments includes 68 subjects and 1900 dialog utterances containing 9200 words. The total recording time amounts to 2 hours and 16 minutes. The most frequent 200 words out of the total of 720 words cover 92 percent of the collected transcribed data. About 700 utterances are simple answers to system questions while the rest, 1200, can be regarded as user initiatives.

We can find a few examples of restarts in the database due to hesitations or mistakes on the semantic, grammatical or phonetic level. However, less than 3% of the utterances contain such disfluencies. Some of the restarts are exact repetitions of a word or a phrase. In some cases a preposition, a question word or a content word is changed. The average utterance length was 5.6 words. The average length of the first sentence in each scenario was 8.8 words.

4.2. Parser evaluation

The parser has been evaluated in several different ways. Most tests used a deleted estimation procedure. Using about 1700 sentences in the Waxholm database, 62 percent give a complete parse, whereas if we restrict the data to utterances containing user initiatives (about 1200), the result is reduced to 48 percent. This can be explained by the fact that the large number of responses to system questions typically have a very simple syntax.

If we exclude extralinguistic sounds such as lip smack, sigh and laughing in the test material based on dialog initiatives by the user, we increase the result to 60 percent complete parses. Sentences with incomplete parses are handled by the robust parsing component and frequently effect the desired system response.

The perplexity on the Waxholm material is about 26 using a trained grammar. If only utterances with complete parses are considered we get a perplexity of 23.

4.3. N-best resorting

The parser has also been evaluated in an N-best list resorting framework. Totally 290 N-best lists with about 10 alternatives each were generated, using an early version of the speech recognition module of the Waxholm system [12]. Since several of the utterances were answers to simple questions the utterance length only averaged about 5 words. The top choice using a bigram grammar as part of the recognition module gave a word accuracy of 76.0%. The mean worst and best possible accuracy in the lists were 48.0% and 86.1%. After resorting using the STINA parser the result improved to 78.6% corresponding to about 25% of the possible increase.

4.4. Evaluation of topic selection

We have performed a sequence of tests to evaluate the topic selection method. The evaluation has used one quarter of the material, about 300 utterances, as test material, and the rest as training material, about 900 utterances. This procedure has been repeated for all quarters and the reported results are the mean values from these four runs. The first result, 12.9% errors in Table 1, is based on the unprocessed labeled input transcription. The eight possible topics have a rather uneven distribution in the material with TIME_TABLE occurring 45% of the time. One of the topics, labeled "no understanding," is trained on a set of constructed utterances that are not possible to understand, even for a human. This topic is then used as a model for the system to give an appropriate "no understanding" system response. It should be noted that, in principle, this is not a question of

utterances that do not get a reasonable parse. However, the topic prediction is certainly influenced by this fact. It seemed reasonable to exclude the "no understanding" prediction from the result since the system at least does not make an erroneous decision. The accuracy model in word recognition evaluation has the same underlying principle. By excluding 55 utterances, about 5% of the test corpus, predicted to be part of the "no understanding" topic, we reduce the error by about 4%.

In the next experiment, we excluded all extralinguistic sounds, about 700, in the input text. This will increase the number of complete parses with about 10% as discussed earlier. The prediction result was about the same compared to the first experiment.

The final experiment included only those utterances that gave a complete parse in the analysis. The errors were drastically reduced. We do not yet know if an increased grammatical coverage also will reduce the topic prediction errors.

All topics		
Test material	N	% Error
woz input	1209	12.9
no extralinguistic sounds	1214	12.7
only complete parses	581	3.1
All topics excluding no "understanding"		
Test material	N	% Error
woz input	1154	8.8
no extralinguistic sounds	1159	8.5
only complete parses	580	2.9

Table 1. Results from the topic prediction experiments.

5. Summary

Lexical semantic information combined with the grammar rules describe the system constraints in our system. Thus, the choice of semantic features and terminal nodes will automatically turn the general grammar into a subgrammar based on the domain. The use of N-gram statistics improves the predictive power of the grammar on both terminal level and phrase structure. Topic prediction based on semantic features separates the surface form of an utterance from the intention of the subject. The dialog design can be data driven to some extent with the proposed method. The rule-based, and to some extent, probabilistic approach we are exploring makes the addition of new topics relatively easy. However, much manual work still remains to be done when an application domain should be changed.

6. Acknowledgment

This work has been supported by grants from The Swedish National Language Technology Program.

7. References

1. Bertenstam J, Blomberg M, Carlson R, Elenius K, Granström B, Gustafson J, Hunnicutt S, Högberg J, Lindell R, Neovius L, de Serpa-Leitao A, Nord L and Ström N. "The Waxholm Application Data-Base." Proc Eurospeech '95, Madrid, 833-836, 1995.

2. Blomberg M, Carlson R, Elenius K, Granström B, Gustafson J, Hunnicutt S, Lindell R and Neovius L. "An experimental dialog system: Waxholm." Proc Eurospeech '93, Berlin, 1867-1870, 1993.

3. Carlson R and Hunnicutt S. "The natural language component - STINA." STL-QPSR, KTH, 1: 29-48, 1995.

4. Carlson R, Hunnicutt S and Gustafson J. "Dialog management in the Waxholm system." Proc. ESCA /ETRW on Spoken Dialog Systems, Vigsø, 137-140, 1995.

5. Seneff S (1992). "TINA: A natural language system for spoken language applications." Computational Linguistics, 18:1:61-66.

6. Jelinek F, Mercer R., and Roukos R. "Principles of lexical language modeling for speech recognition." In Advances in Speech Signal Processing Eds. Furui S and Sondhi M. Marcel Dekker, 651-699, 1992.

7. Stolcke A. "Combining N-grams and SCFGs in Speech Language Models." In proc IEEE Automatic Speech Recognition Workshop, Snowbird, USA, 1995.

8. Seneff S, McCandless M and Zue V. "Integrating Natural Language into the Word Graph Search for Simultaneous Speech Recognition and Understanding." Proc Eurospeech '95, Madrid, 1781-1784, 1995.

9. Moore R, Appelt D, Dowlding J, Gawon M and Moran D. "Combining Linguistic and Statistical Knowledge Sources in Natural-Language Processing for ATIS." In Proc ARPA Workshop on Human Language Technology, 1995.

10. Gorin A. "Semantic associations, acoustic metrics and adaptive language aquisition." Proc ICSLP, Yokohama, 79-82, 1994.

11. Nowel P and Moore R. "The application of dynamic programming techniques to non-word based topic spotting." Proc Eurospeech '95, Madrid, 1355-1358, 1995.

12. Ström N. "Optimising the lexical representation to improve A* lexical search." STL-QPSR, KTH, 2-3: 113-124, 1994.